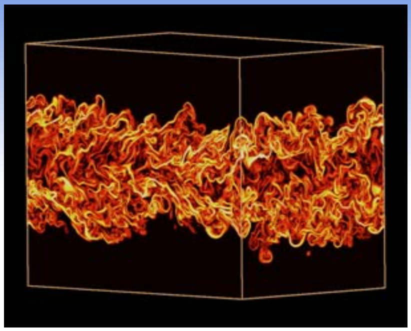
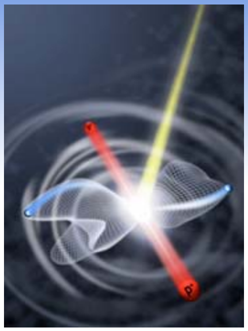
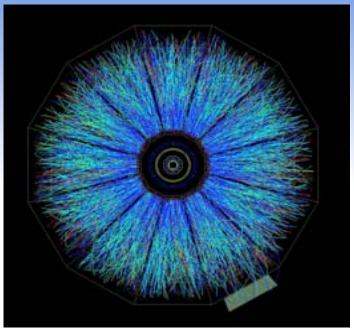
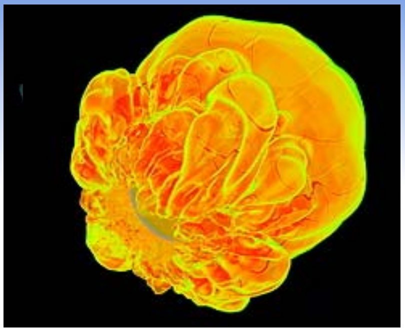






NERSC Role in Biological and Environmental Research

Katherine Yelick
NERSC Director

Requirements Workshop

NERSC is the Production Facility for DOE SC

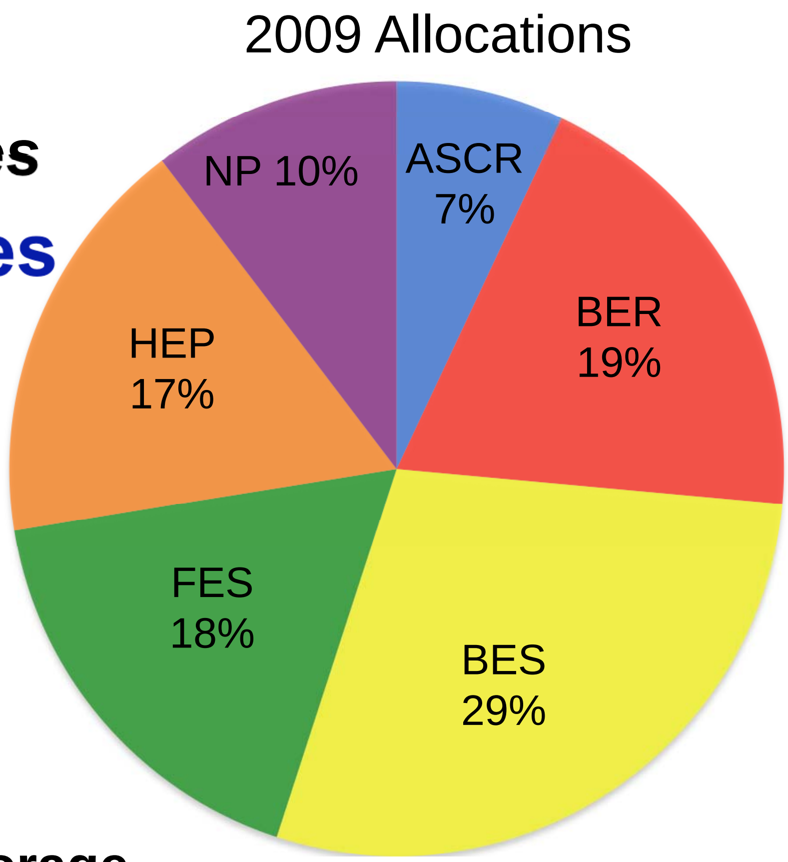
- **NERSC serves a large population**

  Approximately 3000 users,

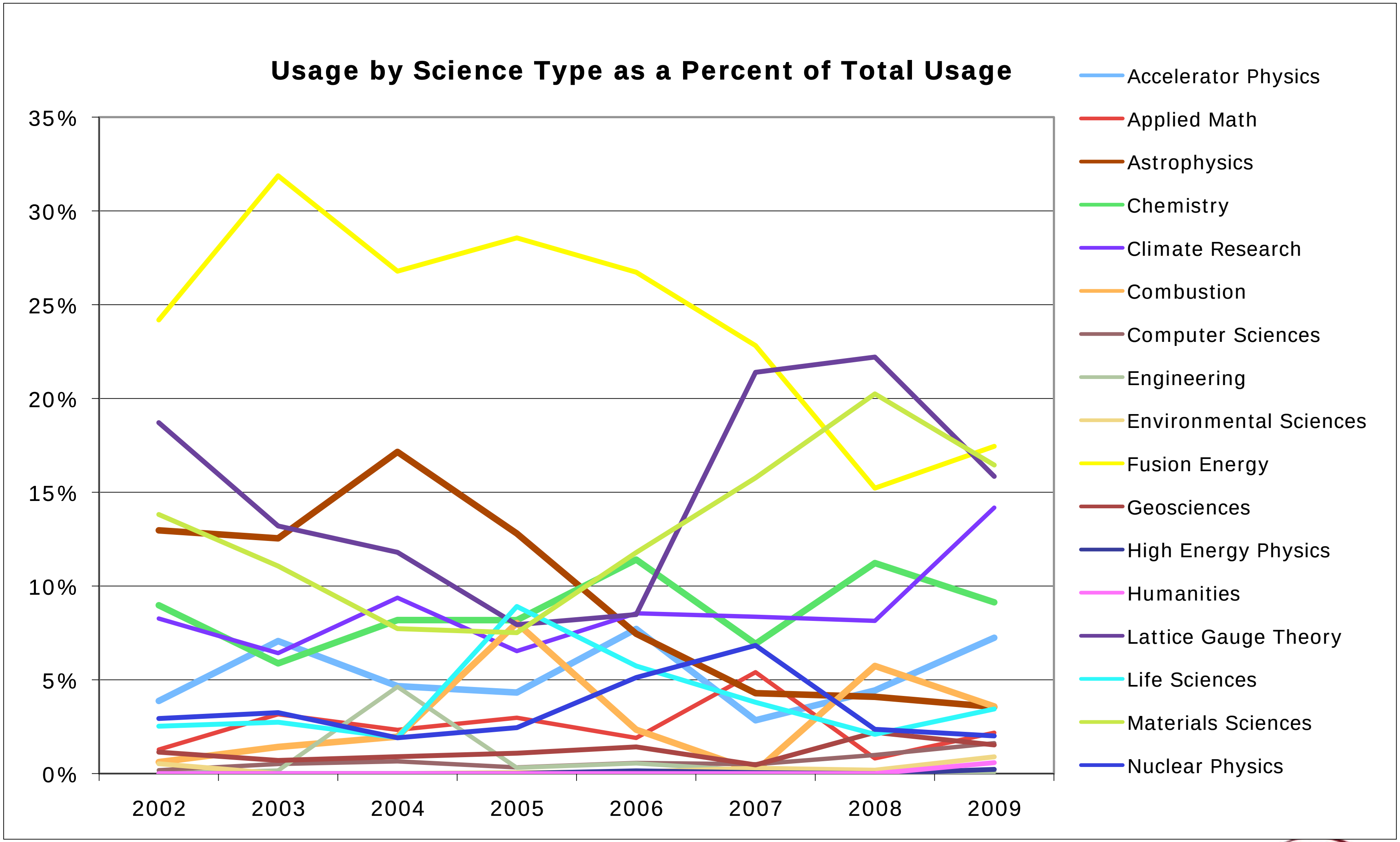
  400 projects, 500 code instances

- **Focus on "unique" resources**
  - **High end computing systems**
  - **High end storage systems**
    - **Large shared file system**
    - **Tape archive**
  - **Interface to high speed networking**
    - **ESNEt soon to be 100 Gb/s**

- **Allocate time / storage**
  - **Current processor hours and tape storage**



What's Changed in DOE Priorities for NERSC?



ASCR's Computing Facilities

NERSC

LBNL

- Hundreds of projects
- 2010 allocations:
  - 70-80% SC offices control; ERCAP process
  - 10-20% ASCR (new ALCC program)
  - 10% NERSC reserve
- Science covers all of DOE/SC science

LCFs

ORNL and ANL

- Tens of projects
- 2010 allocations:
  - 70-80% ANL/ORNL managed; INCITE process
  - 10-20% ACSR (new ALCC program)
  - 10% LCF reserve
- Science areas limited to those at largest scale; not limited to DOE/SC

NERSC 2009 Configuration

Large-Scale Computing System

Franklin (NERSC-5): Cray XT4

- 9,532 compute nodes; 38,128 cores
- ~25 Tflops/s sustained application performance
- 356 Tflops/s peak performance
- 8 GB of memory per quad-core node

Clusters

Bassi (NCSb)

- IBM Power5 (888 cores)

Jacquard (NCSa)

- LNXI Opteron (712 cores)

PDSF (HEP/NP)

- Linux cluster (~1K cores)

NERSC Global Filesystem (NGF)

Uses IBM's GPFS

440 TB; 5.5 GB/s

HPSS Archival Storage

- 59 PB capacity
- 11 Tape libraries
- 140 TB disk cache

Analytics / Visualization

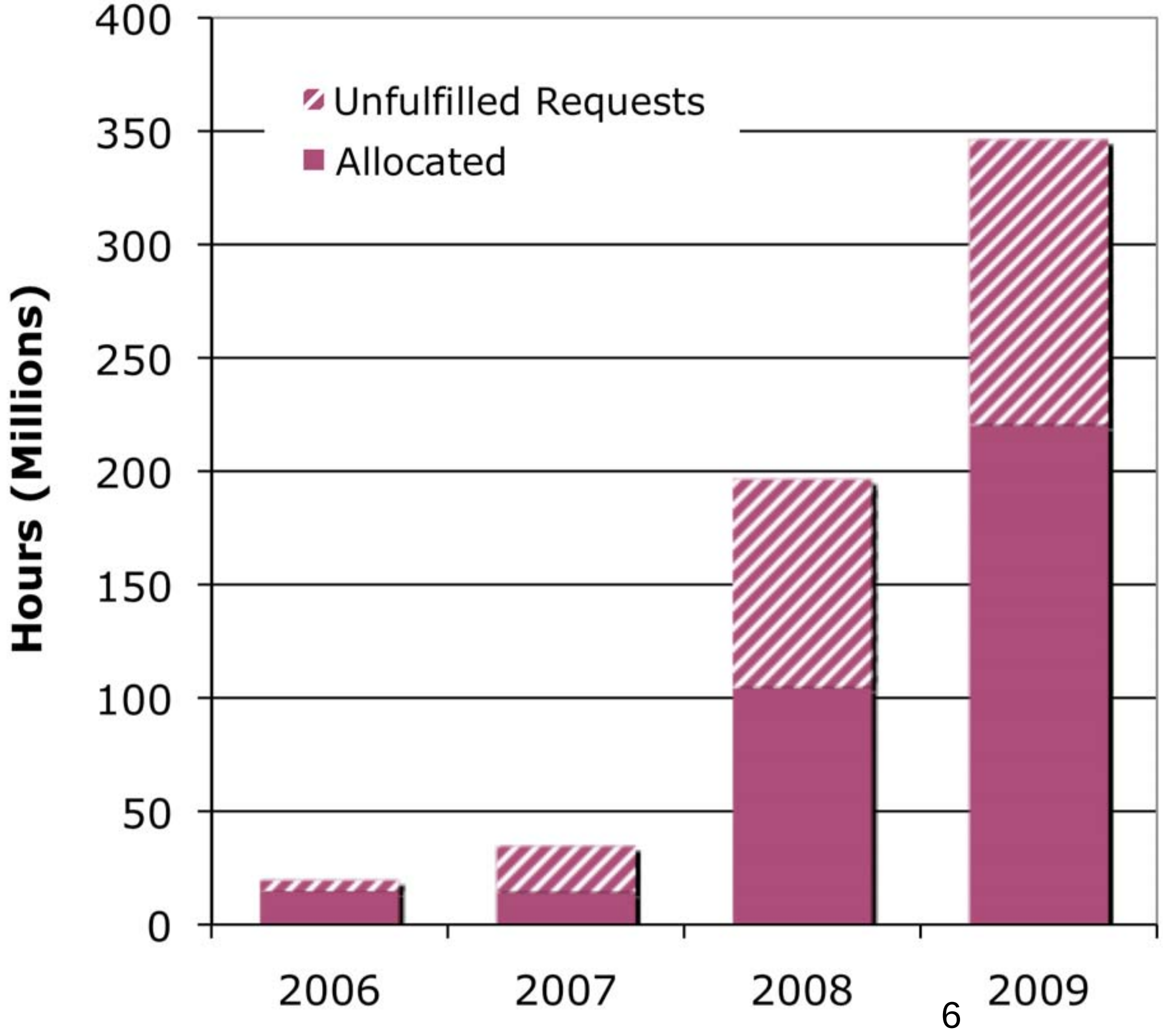
- Davinci (SGI Altix)

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Demand for More Computing

Compute Hours Requested vs Allocated

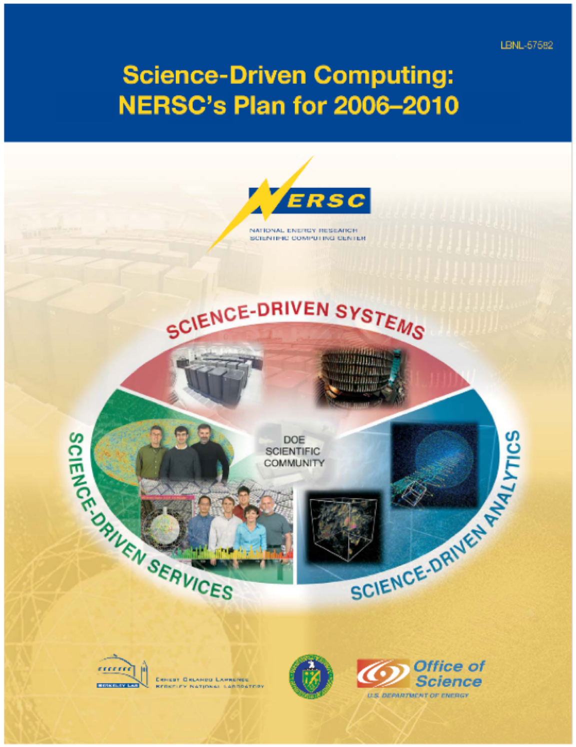


- *Each year DOE users requests ~2x as many hours as can be allocated*
- *This 2x is artificially constrained by perceived availability*
- *Unfulfilled allocation requests amount to hundreds of millions of compute hours in 2009*

How NERSC Uses Your Requirements

2005: NERSC Five-Year Plan

- **Trends:**
  - Widening gap between application performance and peak
  - Emergence of multidisciplinary teams
  - Flood of scientific data from simulations and experiments
- **NERSC Five-Year Plan**
  - New major system every 3 years, each runs for 5-6 years
- **Implementation**
  - NERSC-5 (Franklin) and NERSC-6 (underway)
  - Clusters (Jacquard, Bassi, TBD) and Davinci



– **Question: What trends do you see in algorithms and usage?**

Applications Drive NERSC Procurements

Because hardware peak performance does not necessarily reflect real application performance

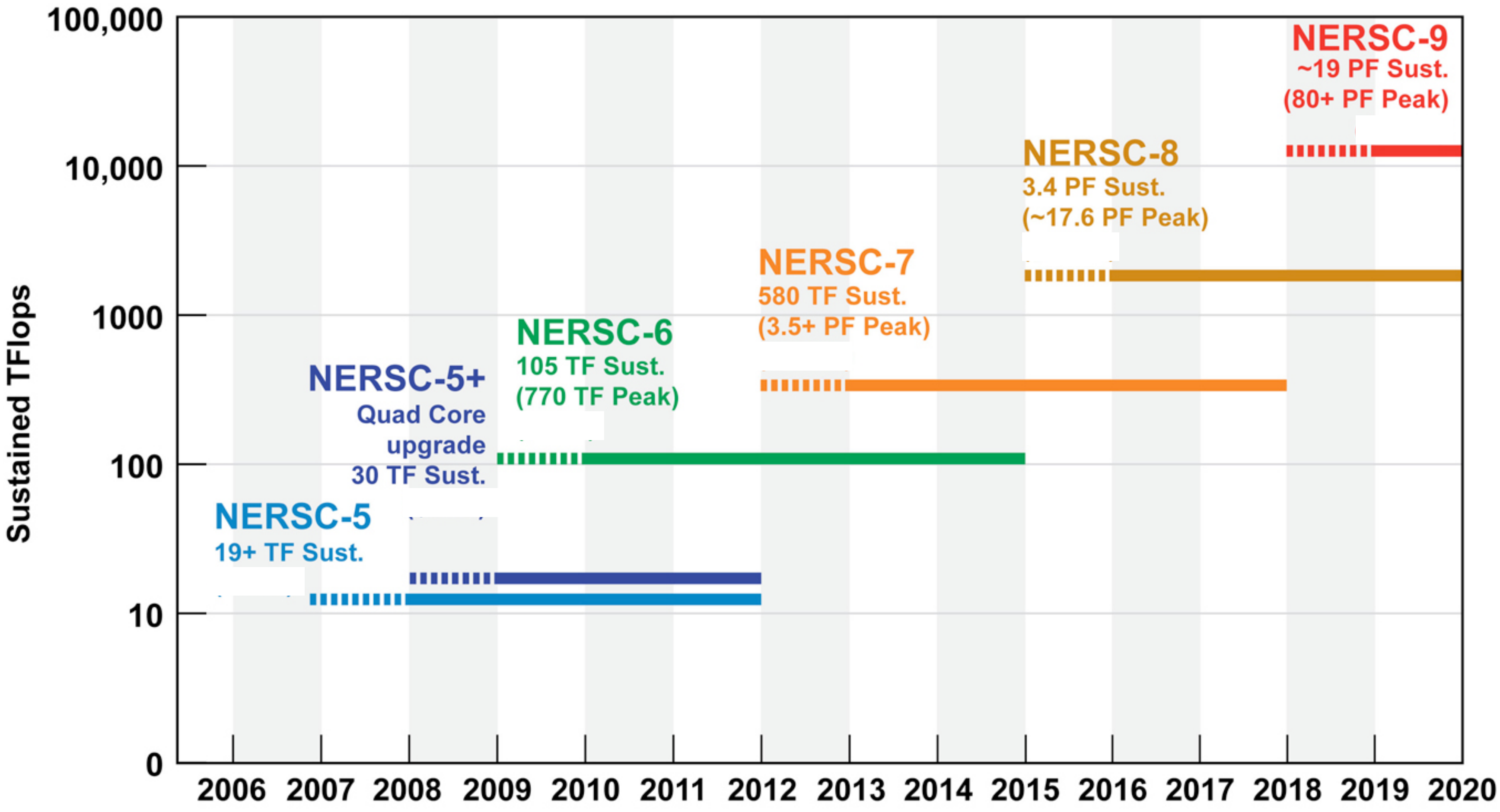
NERSC-6 “SSP” Benchmarks

CAM	*GAMESS*	*GTC*	*IMPACT-T*	*MAESTRO*	*MILC*	*PARATEC*
Climate	Quantum Chemistry	Fusion	Accelerator Physics	Astro-physics	Nuclear Physics	Material Science

- Benchmarks reflect diversity of science and algorithms
- SSP = average performance (Tflops/sec) across machine
- Used before selection, during and after installation
- Question: What applications best reflect your workload?

Requirements Drive NERSC's Long-Term Vision

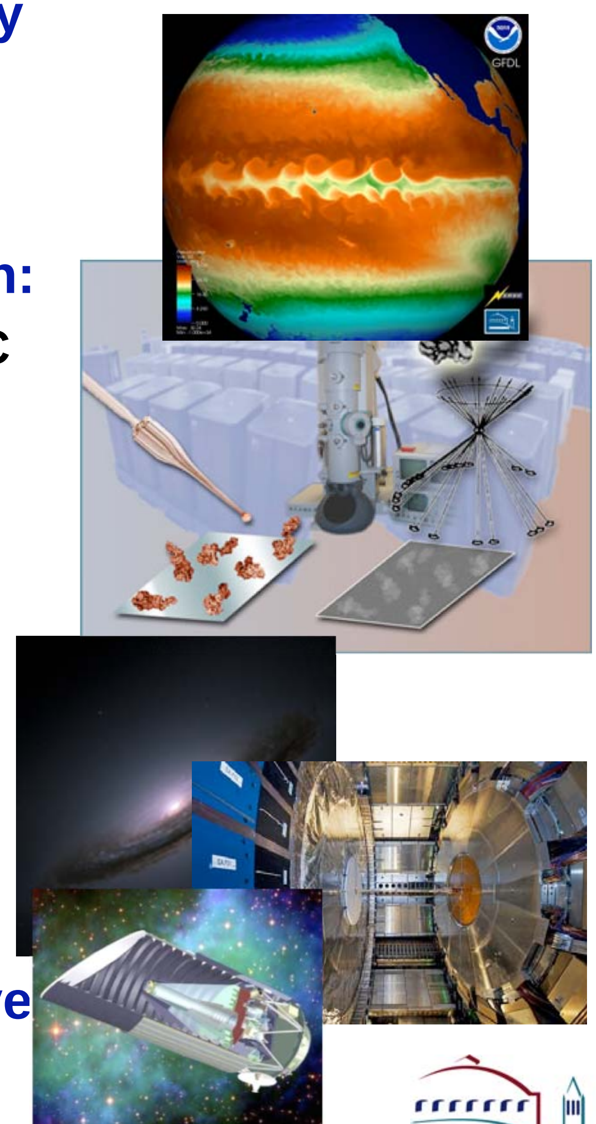
NERSC Plans Circa 2007



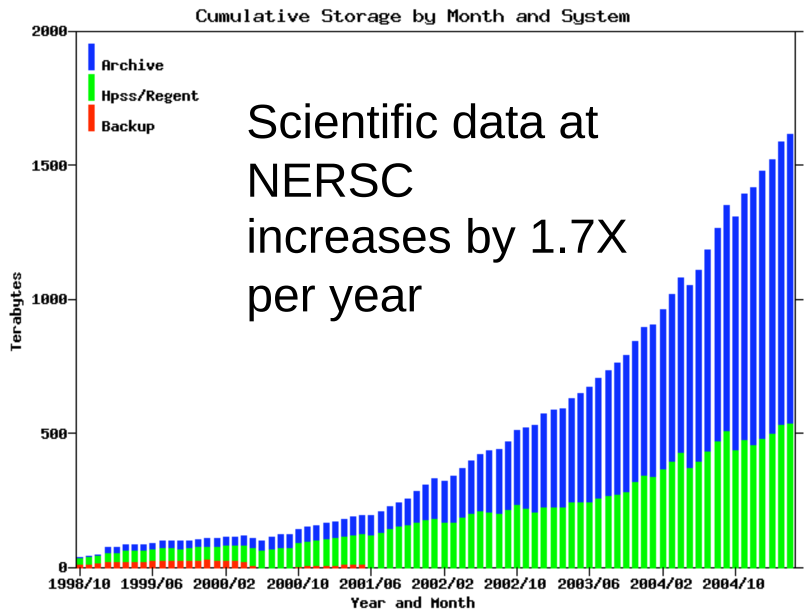
Question: Where do your discoveries lie on this graph?

Data Needs Continue to Grow

- **Scientific data sets are growing exponentially**
  - **Simulation systems and some experimental and observational devices grow in capability with Moore's Law**
- **Petabyte (PB) data sets will soon be common:**
  - ***Climate modeling:* estimates of the next IPCC data is in 10s of petabytes**
  - ***Genome:* JGI alone will have .5 petabyte of data this year and double each year**
  - ***Particle physics*: LHC are projected to produce 16 petabytes of data per year**
  - ***Astrophysics*: JDEM alone will produce .7 petabytes/year**
- **We will soon have more data than we can effectively store and analyze**
- **Question: What are your data set sizes (active disk vs. archive), bandwidths?**

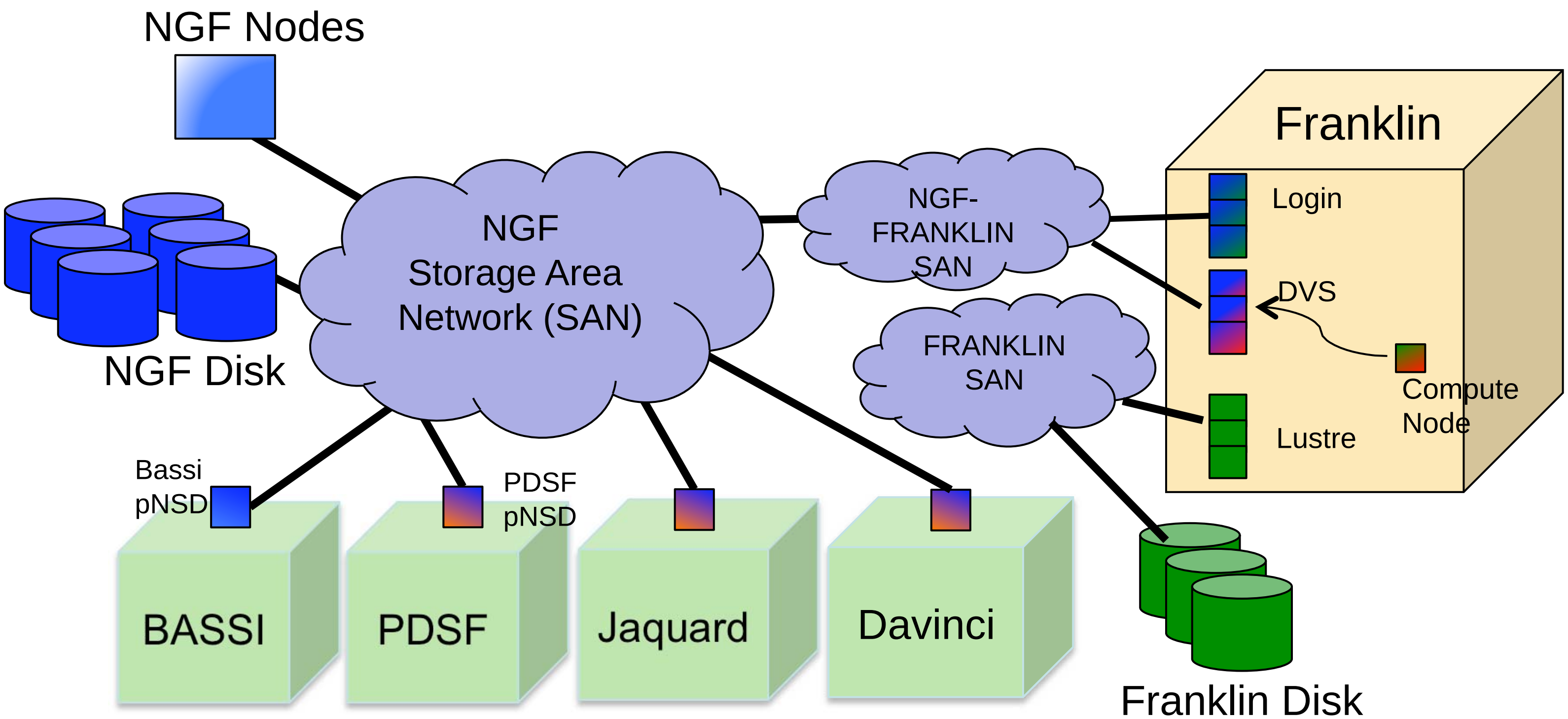


Tape Archives: Green Storage



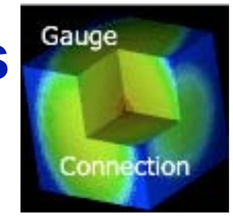
- **Tape archives are important to efficient science**
  - **2-3 orders of magnitude less power than disk**
  - **Requires specialized staff and major capital investment**
  - **NERSC participates in development (HPSS consortium)**
- **Questions: What are your data sets sizes and growth rates?**

NERSC Global File system (NGF)



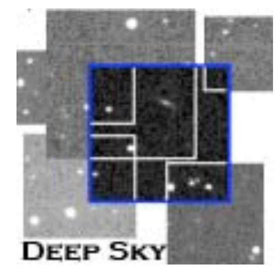
- A facility-wide, high performance, parallel file system
  - Uses IBM's GPFS technology for scalable high performance
  - Makes users more productive
- Questions: How large is your "working set"? Is it shared community-wide?

Science Gateways



- **Create scientific communities around data sets**
  - **Models for sharing vs. privacy differ across communities**
  - **Accessible by broad community for exploration, scientific discovery, and validation of results**
  - **Value of data also varies: observations may be irreplaceable**

- **A *science gateway* is a set of hardware and software that provides data/services remotely**
  - **Deep Sky – "Google-Maps" of astronomical image data**
    - **Discovered 36 supernovae in 6 nights during the PTF Survey**
    - **15 collaborators worldwide worked for 24 hours non-stop**
  - **GCRM – Interactive subselection of climate data**
  - **Gauge Connection – Access QCD Lattice data sets**
  - **Planck Portal – Access to Planck Data**



- **Building blocks for science on the web**
  - **Remote data analysis, databases, job submission**

Traditional Sources of Performance Improvement are Flat-Lining

- **Moore's Law is alive and well**
- **15 years of *exponential* clock speed growth has ended**
- **How to use the transistors?**
  - Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!
  - *Is this a good idea, or is it completely brain-dead?*
  - *Concurrency will increase, but how much from SIMD, vectors, cores, accelerators?*

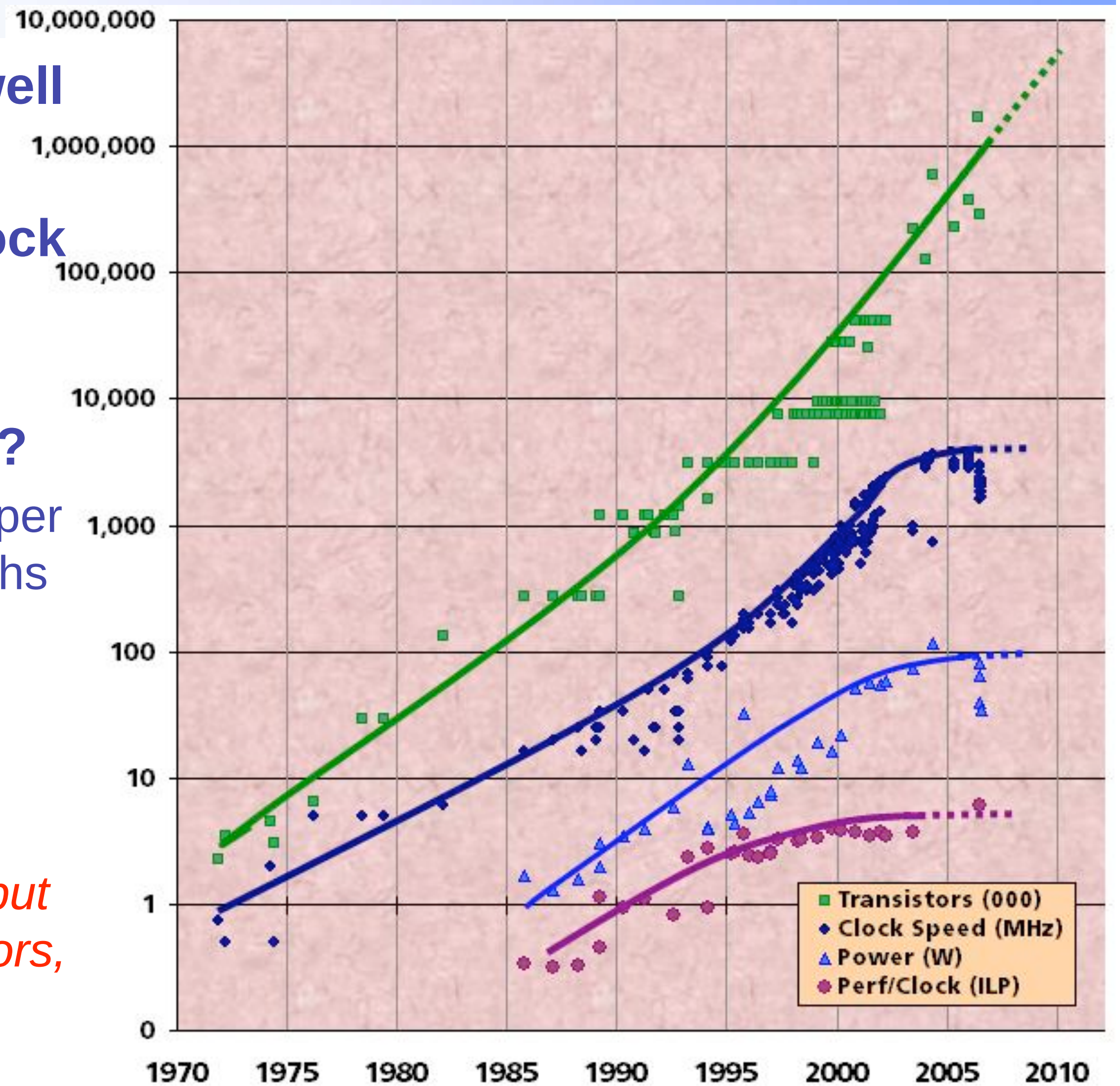
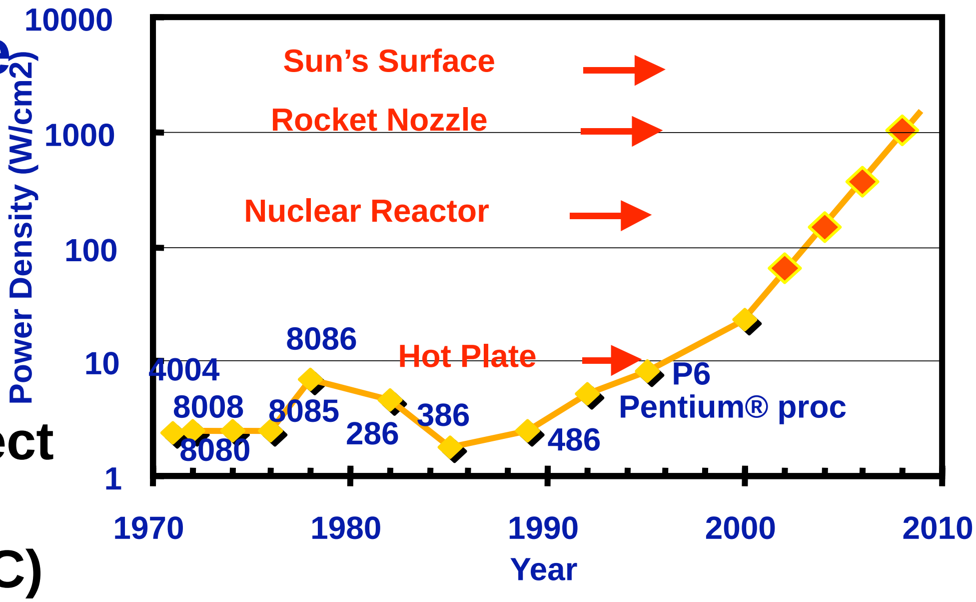


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Parallelism is "Green"

- **Concurrent systems are more power efficient**
  - Dynamic power is proportional to V^2fC
  - Increasing frequency (f) also increases supply voltage (V) → cubic effect
  - Increasing cores increases capacitance (C) but only linearly

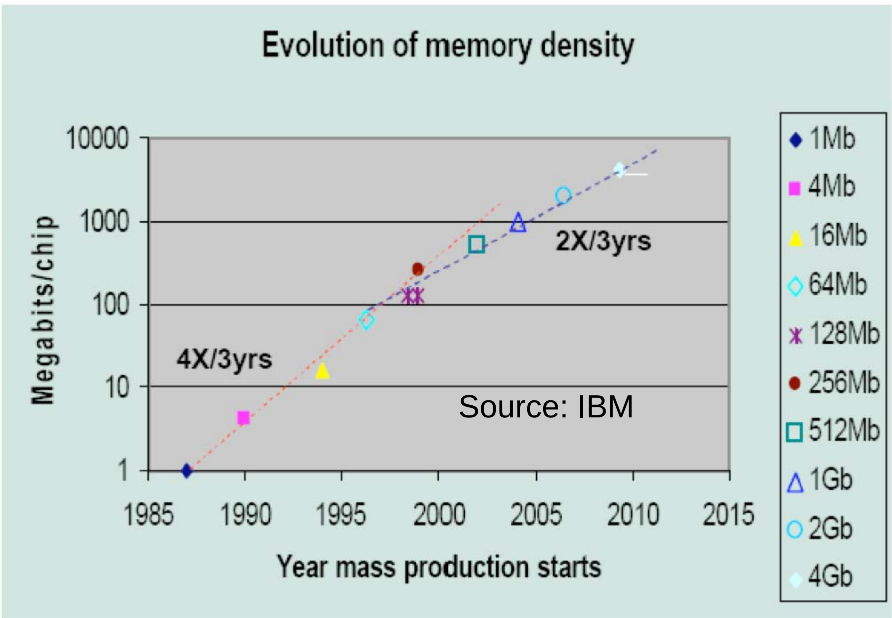


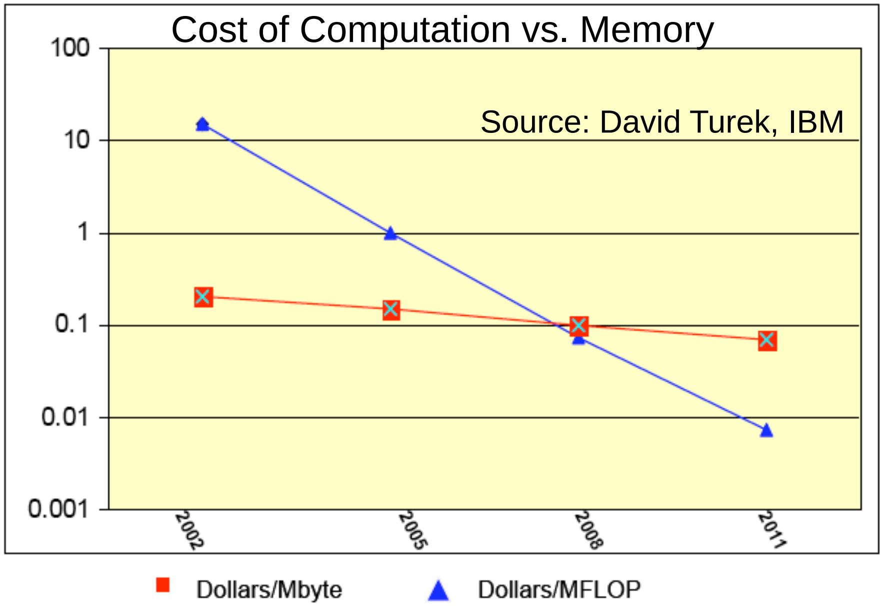
- **High performance serial processors waste power**
  - Speculation, dynamic dependence checking, etc. burn power
  - Implicit parallelism discovery
- **Question: *Can you double the concurrency in your algorithms and software every 2 years?***

Technology Challenge

Technology trends against a constant or increasing memory per core

- Memory density is doubling every three years; processor logic is every two
- Storage costs (dollars/Mbyte) are dropping gradually compared to logic costs





The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it

Question: *Can you double concurrency without doubling memory?*

Hardware and Software Trends

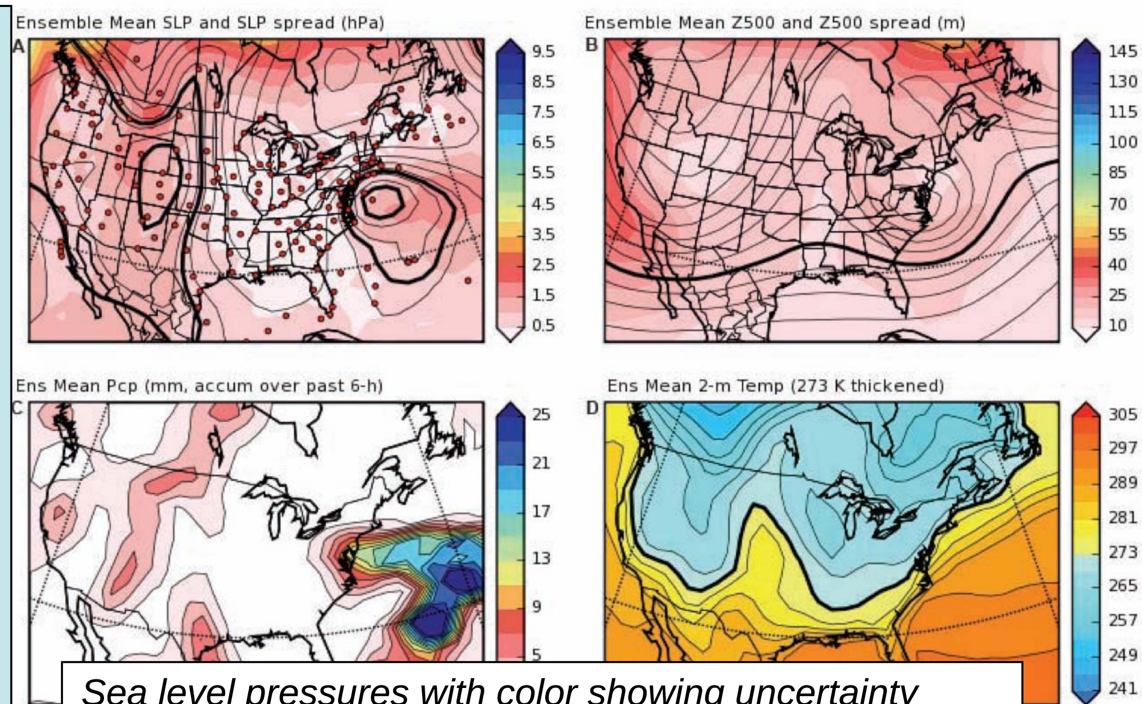
- **Hardware Trends**
  - Exponential growth in explicit on-chip parallelism
  - Reduced memory per core
  - Heterogeneous computing platforms (e.g., GPUs)
  - As always, this is largely driven by non HPC markets
- **Software Response**
  - Need to express more explicit parallelism
  - New programming models on chip: MPI + X
  - Increased emphasis on strong scaling
- **What we want**
  - Understand your requirements and help craft a strategy for transitioning to a hardware and programming environment solution

Biological and Environmental Science at NERSC

Validating Weather Models

- **"20th Century Reanalysis" using an Ensemble Kalman filter to fill in missing climate data since; can be used for validation of models**
- **PI: G. Compo, U. Boulder**

- Science Results:
  - Reproduced 1922 Knickerbocker storm and dust storms of 1930s
  - Building maps every 6 hours 1982-2008
- Scaling Results:
  - Scales to 2.4K cores
  - Switched to higher resolution algorithm with Franklin access
  - 1M hours in 2009 (NERSC Reserve)



Sea level pressures with color showing uncertainty (a&b); precipitation (c); temperature (d). Dots indicate measurements locations (a).

Supporting Efficient Algorithms

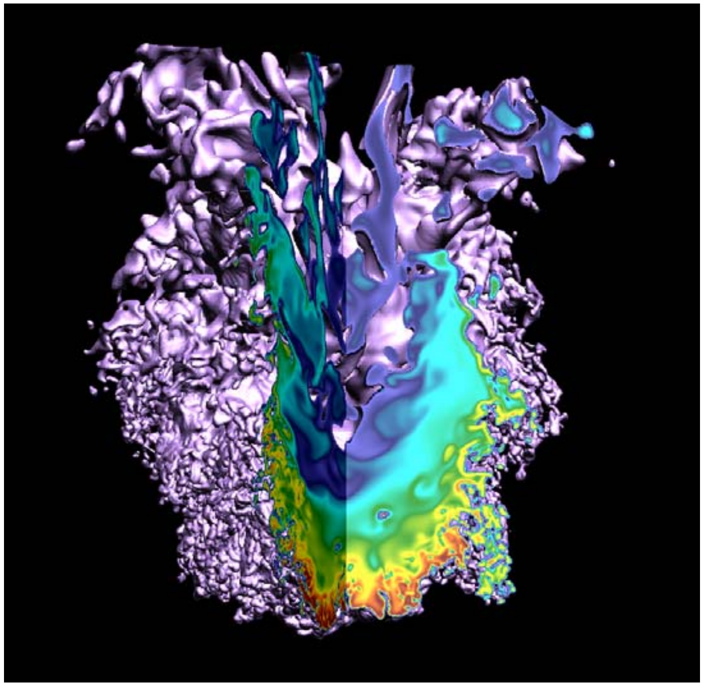
- **Computation: Numerical simulation of a lean premixed hydrogen flame in a laboratory-scale low-swirl burner (LMC code). Uses a low Mach number formulation, adaptive mesh refinement (AMR) and detailed chemistry and transport.**
- **PI: John Bell, LBNL**

Science Result:

- Simulations capture cellular structure of lean hydrogen flames and provide a quantitative characterization of enhanced local burning structure

NERSC Results:

- LMC dramatically reduces time and memory.
- Scales to 4K cores, typically run at 2K
- Used 9.6M hours in 2008, allocated 5.5M in 2009

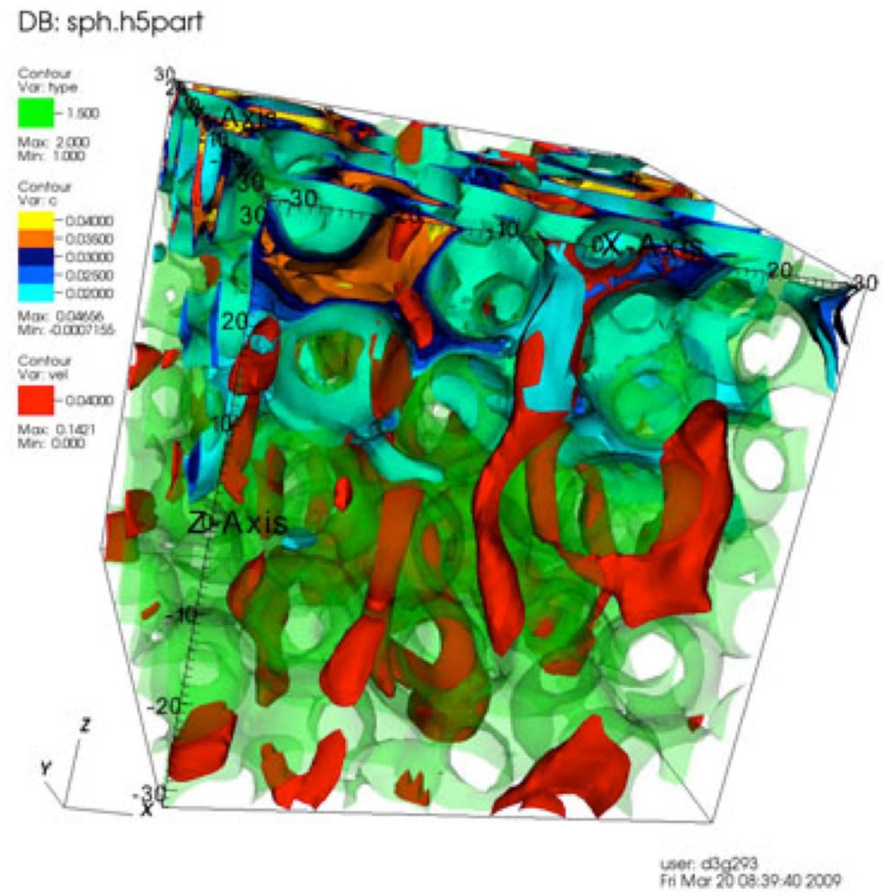


J B Bell, R K Cheng, M S Day, V E Beckner and M J Lijewski, Journal of Physics: Conference Series 125 (2008) 012027

Supporting Efficient Algorithms

- **Hybrid numerical methods for subsurface Biogeochemical processes**
- **PIs: Tim Sheibe, Bruce Palmer, et al**

NERSC Analytics collaborating on
  Data Model for particle data
  Parallel Visualization (VisIt)



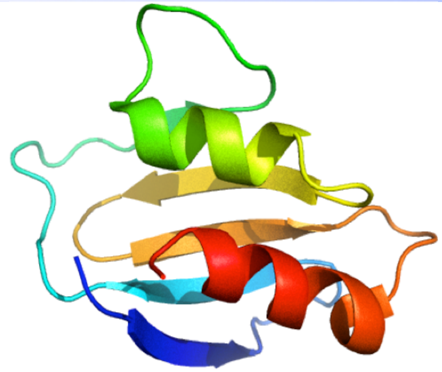
NERSC Service Examples: Molecular Dynamics and Protein Folds

PI: Valerie Dagget

Science Goals:

Catalog dynamical shapes of proteins by systematically unfolding them.

Results include increased sampling of biomedically relevant targets

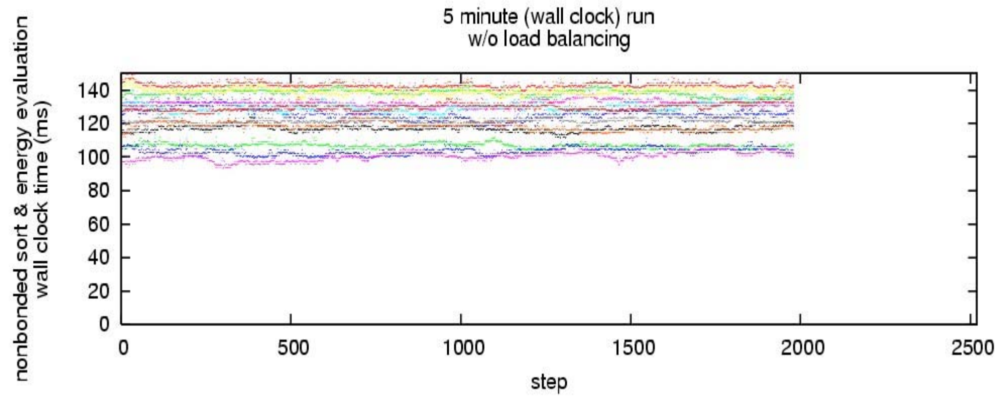


Work with NERSC:

Load balance in the *ilmm* code

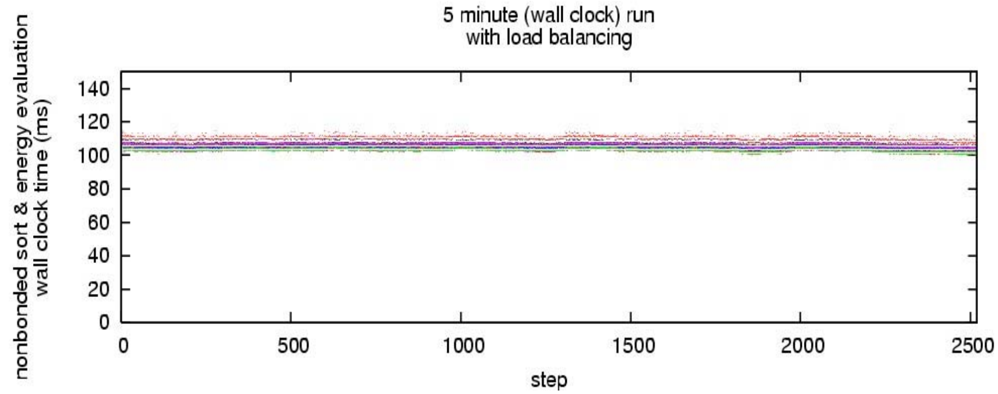
Scalar optimizations

Batch work flow planning



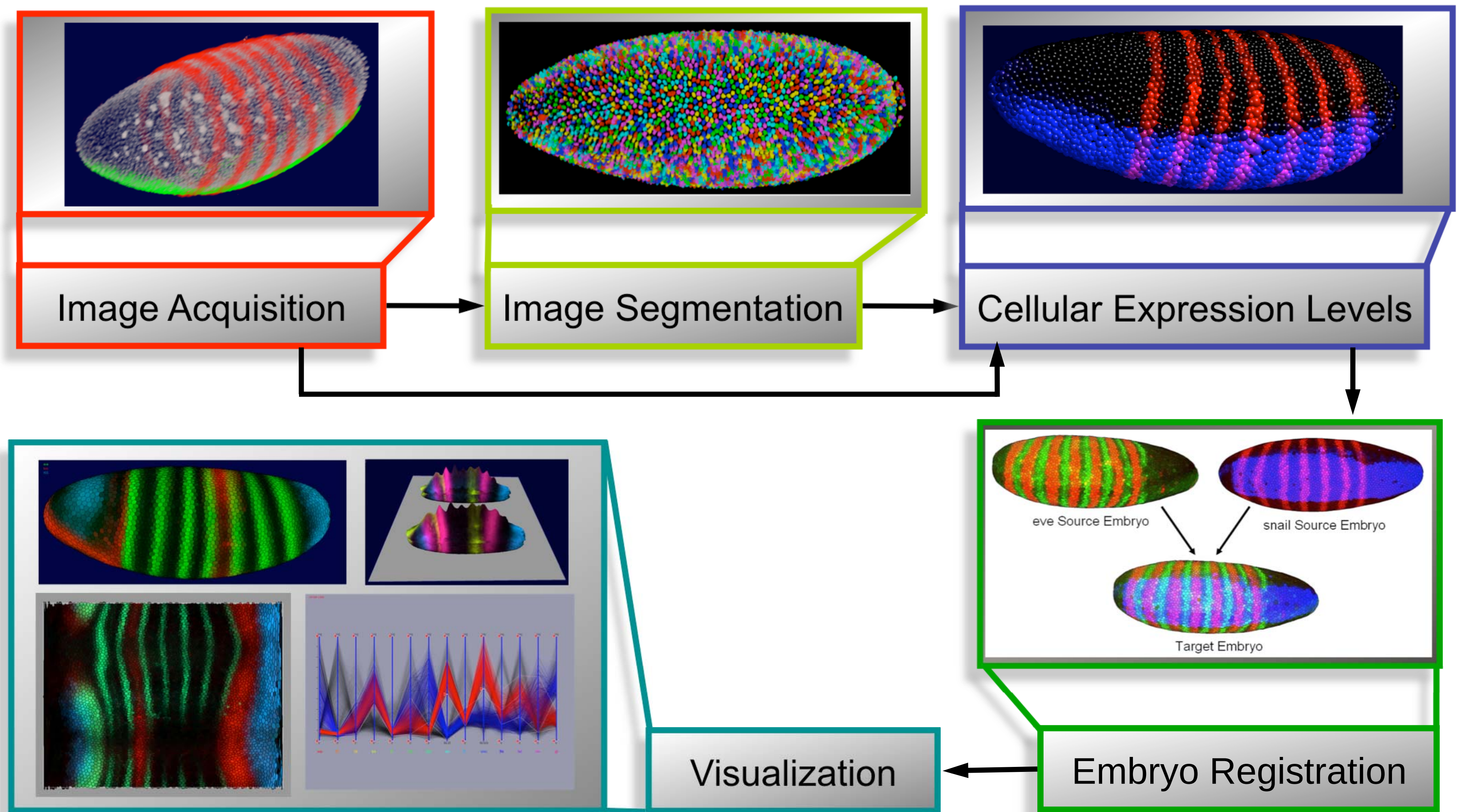
Impact:

20% improvement in time to solution. Portable code improvements.



U.S. DEPARTMENT OF ENERGY | Office of Science

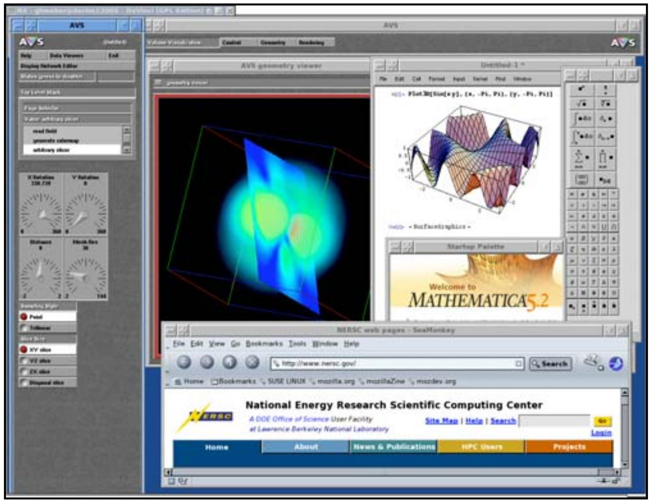
Data Acquisition and Visualization Pipeline



NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Accelerating Remote Display

- **Problem: remote display operations are very slow due to network latency.**
- **Solution: deploy new technology at NERSC that hides network latency in remote display operations to improve user productivity.**
- **Deployed Summer 2008 to entire NERSC user community.**
- **Results: improves remote display by a factor of about 10x.**



Screenshot of a remote display session running multiple 3D visual data analysis applications.

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Conclusions

- **NERSC requirements**
  - **Qualitative requirements shape NERSC functionality**
  - **Quantitative requirements set the performance**

    "What gets measure gets improved"

- **Goals:**
  - **Your goal is to make scientific discoveries**
  - **Our goal is to enable you to do science**

Science-Driven Computing Strategy 2006 -2010